Complexity of Gröbner basis computation for Semi-regular Overdetermined sequences over F_2 with solutions in F_2

Magali Bardet, Jean-Charles Faugère, Bruno Salvy

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Complexity of Gröbner basis computation for Semi-regular Overdetermined sequences over $\mathbb{F}_2$ with solutions in $\mathbb{F}_2$

Magali Bardet and Jean-Charles Faugère and Bruno Salvy

N° 5049

Décembre 2003

THÈME 2

ISSN 0249-6399 ISRN INRIA/RR--5049--FR+ENG

LORRAINE

Complexity of Gröbner basis computation for Semi-regular Overdetermined sequences over $\mathbb{F}_2$ with solutions in $\mathbb{F}_2$

Magali Bardet* and Jean-Charles Faugère† and Bruno Salvy‡

Thème 2 — Génie logiciel
et calcul symbolique
Projets Spaces et Algo

Rapport de recherche n° 5049 — Décembre 2003 — 19 pages

Abstract: We present complexity results for solving "typical" overdetermined algebraic systems over $\mathbb{F}_2$ with solutions in $\mathbb{F}_2$, using Gröbner bases. They are useful for instance to predict the complexity of an algebraic cryptanalysis over a cryptosystem, they give *a priori* upper bounds.

We define *semi-regular* sequences and their associated notion of *degree of regularity* D_{reg}. The motivation for studying semi-regular sequences is that "random" sequences are semi-regular, and D_{reg} is closely related to the global cost of the Gröbner basis computation for a graded admissible monomial order. Using in particular the F5 Gröbner basis algorithm, we show that for semi-regular sequences the behavior of F5 (in a matrix version) can be followed step by step, and the size of all matrices made explicit. We deduce D_{reg}, and using asymptotic analysis methods we compute its asymptotic expansion.

We give many explicit examples, and discuss the complexity of the global arithmetic cost of the Gröbner basis computation for $m(n)$ quadratic equations in n variables: for $m \sim Nn$ with N constant, the computation is exponential, if $n \ll m \ll n^2$ the computation is sub-exponential and for $m \sim Nn^2$, with N a constant, the complexity is $\sim \max(n^{2\omega}, n^{\omega/(8N)})$ which is polynomial. This classification gives "generic upper bounds", and thus a priori upper bounds for many cryptosystems.

Key-words: Algebraic cryptanalysis, Gröbner basis, semi-regular sequences, overdetermined algebraic systems

* bardet@calfor.lip6.fr
† jcf@calfor.lip6.fr
‡ Bruno.Salvy@inria.fr

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois, Campus scientifique,
615, rue du Jardin Botanique, BP 101, 54602 Villers-Lès-Nancy (France)
Téléphone : +33 3 83 59 30 00 — Télécopie : +33 3 83 27 83 19

Complexité du calcul de base de Gröbner pour des systèmes semi-réguliers sur-déterminés à coefficients dans $\mathbb{F}_2$ et solutions dans $\mathbb{F}_2$

Résumé : Nous présentons des résultats de complexité pour la résolution par bases de Gröbner de systèmes algébriques sur-déterminés "typiques" à coefficients dans $\mathbb{F}_2$ et solutions dans $\mathbb{F}_2$. Ces résultats sont utiles par exemple pour prédire la complexité d'une cryptanalyse d'un cryptosystème, ils donnent des bornes supérieures "à priori".

Nous définissons les suites semi-régulières et la notion de degré de régularité D_{reg} associée. L'intérêt de l'étude de telles suites est que des suites "aléatoires" sont semi-régulières, et que le degré de régularité est relié au coup global du calcul de base de Gröbner pour un ordre admissible gradué par le degré. En utilisant en particulier l'algorithme F5 de calcul de base de Gröbner, nous montrons que pour des suites semi-régulières le comportement de l'algorithme F5 (version matricielle) peut être suivi pas à pas, et que l'on peut donner explicitement les tailles de toutes les matrices. Nous en déduisons la régularité de telles suites, et en utilisant des méthodes d'analyse asymptotique nous calculons le développement asymptotique de D_{reg}.

Nous donnons de nombreux exemples explicites. Nous discutons du coût global d'un calcul de base de Gröbner pour $m(n)$ équations quadratiques en n variables : pour $m \sim Nn$ avec N constant, le calcul est exponentiel, si $n \ll m \ll n^2$ le calcul est sous-exponentiel et pour $m \sim Nn^2$, avec N constant, la complexitée est $\sim \max(n^{2\omega}, n^{\omega/(8N)})$ qui est polynômiale. Cette classification donne des "bornes supérieures génériques", et donc des bornes supérieures a priori pour de nombreux cryptosystèmes.

Mots-clés : Cryptanalyse algébrique, bases de Gröbner, suites semi-régulières, systèmes algébriques sur-déterminés

1 Introduction

Since the 1980's, many systems appearing in cryptography, e.g. [MI88, Pat96, Moh99, Kob98] have been based on the problem of solving a system of algebraic equations over the finite field $\mathbb{F}_2$, and in many cases the interesting solutions are only solutions in $\mathbb{F}_2$ and not in its algebraic closure. One possibility to find only those solutions is to solve the original system of equations over $\mathbb{F}_2$ together with the field equations $x_1^2 + x_1, \ldots, x_n^2 + x_n$ if the variables are $x_1, \ldots, x_n$. A new general cryptanalysis approach, called algebraic cryptanalysis, appeared more recently. A number of recent papers, e.g. [BDC03, CM03, AK03, MR02, ...] exploit the algebraic structure of the cryptosystem to derive an algebraic system relating for instance the output bits to the input bits. The complexity bounds we give are upper bounds for either the security of the cryptosystem in the first case, or the feasibility of the attack in the second case.

The topic of this paper is the complexity of the resolution of systems of algebraic equations over the finite field $\mathbb{F}_2$ with solutions in $\mathbb{F}_2$ (which means together with the field equations $x_1^2 + x_1, \ldots, x_n^2 + x_n$) using Gröbner bases. Gröbner basis is a well known, generic method for solving algebraic systems, and seems to be very efficient in some cases, for instance when applied to the HFE problem [FJ03]. The study of other specific algorithms like XL [SPCK00] is interesting, but out of the scope of the paper. We conjecture that the bounds we derive in this paper apply to the XL algorithm.

For the *worst case*, a bound on the complexity of the Gröbner basis computation is already known. Thanks to the field equations a system $f_1, \ldots, f_m, x_1^2 + x_1, \ldots, x_n^2 + x_n$ has a finite number of solutions, and also a finite number of solutions at infinity (the solutions at infinity are obtained by homogenizing the equations with a new variable, say x_0, putting $x_0 = 0$ and solving the resulting system). In this case, a result of Lazard [Laz83] applies and the maximal degree of any element of a Gröbner basis for the DRL (Degree Reverse Lexicographic) order is at most the Macaulay bound $n+1$. The global cost of the computation of a Gröbner basis for the system $f_1, \ldots, f_m, x_1^2 + x_1, \ldots, x_n^2 + x_n$ is at most polynomial in 2^n, hence simply exponential in n. The doubly exponential behavior of Gröbner bases computations in the worst case (the Mayr-Meyer example [MM82]) cannot appear in this case.

The problem of solving algebraic equations is known to be a hard problem (NP-complete, even over a finite field, and even if the equations are of degree less than or equal to 2 [FY80]). It is well known that the NP-completeness of a problem is not sufficient for its use in cryptography, the problem has to be hard in average. The goal of this paper is to give complexity results for overdetermined "generic systems", that is for "random" systems (e.g. coefficients are picked at random) or for systems coming from practical problems. Our contributions are the following.

Semi-regular sequences over $\mathbb{F}_2$. First, we give a mathematical definition for "generic systems", called *semi-regular sequences*. We define also the *degree of regularity* of a system, called D_{reg}, and show that this degree of regularity is a bound for the maximal degree of an element of a Gröbner basis, for any degree order (notations, definitions and properties of Gröbner basis are recalled in Section 2). Those definitions of semi-regular sequences and degree of regularity generalize the classical definitions of regular sequences and index of regularity over $\mathbb{Q}$ to systems which are overdetermined, affine and over $\mathbb{F}_2$ with solutions in $\mathbb{F}_2$. We conjecture that the proportion of semi-regular sequences of m polynomials in n variables with degrees $d_1, \ldots, d_m$ tends to 1 as $n \to \infty$. We checked it by computer simulation in many usual cases.

Analysis of Algorithm F5. For semi-regular sequences the behavior of the F5 algorithm [Fau02] by Faugère can be well understood. We use a matrix version of the algorithm, called matrix-F5, and a new criterion to take into account the Frobenius morphism $f^2 + f$. This algorithm constructs matrices incrementally in the degree, such that for semi-regular sequences, all matrices in degree at most D_{reg} are full rank. We deduce from this property a recurrence formula, which gives explicitly the size of all matrices up to degree D_{reg}. This makes explicit the Hilbert series of semi-regular sequence, and gives a method to compute D_{reg}. The maximal degree of any polynomial appearing during the Gröbner basis computation with Algorithm F5 is at most D_{reg}, and for any other Gröbner basis algorithm, this maximal degree is larger than that in F5.

Asymptotic analysis of D_{reg}. An important question from the complexity point of view is the behavior of D_{reg} as n tends to infinity. Indeed, D_{reg} is a bound for the degree of an element in the Gröbner basis, and the global cost of the Gröbner basis computation can be estimated to be the cost of the linear algebra on the largest matrix appearing in Algorithm matrix-F5. We use asymptotic methods to determine the asymptotic expansion of D_{reg}. The first term of the asymptotic expansion is given by the saddle point method, and a full asymptotic expansion can be computed from the coalescent saddle points method, every coefficient being expressed in terms of algebraic numbers and the first root of the Airy function Ai. For instance, for n quadratic equations over $\mathbb{F}_2$, the first four terms are $D_{\mathrm{reg}} = 0.0900\,n + 1.00\,n^{\frac{1}{3}} - 1.58 + O(1/n^{\frac{1}{3}})$ numerically. This asymptotic expansion is a very good approximation of the exact value of D_{reg}, already for $n \geq 3$.

Plan of the paper. After recalling some properties of Gröbner bases in Section 2, we introduce semi-regular sequences in Section 3. Section 4 is devoted to the analysis of Algorithm matrix-F5, which is used to compute the Hilbert series and the degree of regularity D_{reg} of semi-regular sequences. In Section 5, using asymptotic analysis methods we compute the asymptotic expansion of D_{reg} in terms of n. The results are computed explicitly in many usual cases, and a method is given to compute them in any case. As a consequence, we derive in Section 6 a classification of the complexity of the Gröbner basis computation in terms of the number of equations, from the polynomial behavior (when we have $m = Nn^2$ equations in n variables, and N is a constant) to the exponential one (with $m = Nn$ equations, N constant).

2 Solving systems with Gröbner bases

The notion of Gröbner basis was introduced by Buchberger, who gave the first algorithm for their computation [Buc65, Buc70]. The Buchberger algorithm is implemented in all Computer Algebra Systems (e.g. Maple, Magma, Cocoa, Singular, Macaulay, Gb, ...).

In this section, we recall basic definitions of Gröbner bases, give some useful properties, in particular over the finite field $\mathbb{F}_2$, and recall the link between Gröbner bases and linear algebra.

2.1 Properties of Gröbner bases

We refer to [CLO97] for details on Gröbner bases (see also [Kob98]). Denote by $S_n = \mathbb{F}_2[x_1, \ldots, x_n]$ the polynomial ring. Let $f_1, \ldots, f_m \in S_n$ be a set of m equations, and let $I = \langle f_1, \ldots, f_m \rangle$ be the ideal of S_n generated by the f_i's. We choose $<$ an admissible monomial order on the set of all monomials in $x_1, \ldots, x_n$.

Example 2.1 *The Degree Reverse Lexicographic order (DRL) defined by* $m_1 = x_1^{\alpha_1} \dots x_n^{\alpha_n} <_{DRL} m_2 = x_1^{\beta_1} \dots x_n^{\beta_n}$ *if degree*$(m_1) <$ *degree*(m_2) *or degree*$(m_1) =$ *degree*(m_2) *and in the n-tuple* $(\alpha_1 - \beta_1, \dots, \alpha_n - \beta_n)$ *the right-most nonzero entry is positive is an admissible monomial order.*

The DRL order belongs to the class of *total degree orders*: one first compares total degrees, and then break ties with some other order. The DRL order is in general the best one for the complexity point of view, and it is also true for any total degree order for systems with solutions in $\mathbb{F}_2$.

Let $LT(f)$ be the leading term of a polynomial f w.r.t. $<$, and $LT(F) = \{LT(f) : f \in F\}$.

Definition 1 (Gröbner basis) *A finite subset* $G \subset I$ *is a Gröbner basis of* I *if for any polynomial* $f \in S_n$ *we have* $f \in I \Rightarrow \exists g \in G\ LT(g) | LT(f)$*, or equivalently if* $\langle LT(G) \rangle = \langle LT(I) \rangle$.

Definition 2 (Reduced Gröbner basis) *A finite subset* $G \subset I$ *is a reduced Gröbner basis of* I *if for all* $g \in G$*, no monomial of* g *lies in* $\langle LT(G \backslash \{g\}) \rangle$.

If $\mathbb{L} \supset \mathbb{F}_2$ is a field extension, the algebraic variety associated to $F = \{f_1, \dots, f_m\}$ over $\mathbb{L}$ is the set of solutions of F in $\mathbb{L}$,

$$V_{\mathbb{L}}(f_1, \dots, f_m) = \{(z_1, \dots, x_n) \in \mathbb{L} : f_i(z_1, \dots, z_n) = 0\ \forall i = 1 \dots m\}$$

A Gröbner basis of I describes all the solutions $V_{\overline{\mathbb{F}_2}}(I)$ of I over $\overline{\mathbb{F}_2}$, where $\overline{\mathbb{F}_2}$ is the algebraic closure of the field $\mathbb{F}_2$.

For cryptographic applications, we want to compute the solutions of a system $F \subset \mathbb{F}_2[x_1, \dots, x_n]$ in the field $\mathbb{F}_2$, and not in the algebraic closure of $\mathbb{F}_2$. In this case, we can use the relation: $V_{\mathbb{F}_2}(\langle f_1, \dots, f_m \rangle) = V_{\overline{\mathbb{F}_2}}(\langle f_1, \dots, f_m, x_1^2 + x_1, \dots, x_n^2 + x_n \rangle)$, which means that if we add to I the field equations over $\mathbb{F}_2$ then the Gröbner basis describes exactly the solutions of the system in $\mathbb{F}_2^n$. In all the following, by computing a Gröbner basis of $\{f_1, \dots, f_m\}$ over $\mathbb{F}_2$ with solutions in $\mathbb{F}_2$ we mean computing a Gröbner basis of $\{x_1^2+x_1, \dots, x_n^2+x_n, f_1, \dots, f_m\}$ and abbreviate it in computing a Gröbner basis of $f_1, \dots, f_m$ over $\mathbb{F}_2$. We denote by $R_n = \mathbb{F}_2[x_1, \dots, x_n]/(x_1^2 + x_1, \dots, x_n^2 + x_n)$ the polynomial ring where any monomial is square-free (reduced modulo the field equations), then finding a Gröbner basis of $\{x_1^2+x_1, \dots, x_n^2+x_n, f_1, \dots, f_m\}$ in S_n is equivalent to finding a Gröbner basis of $\{f_1, \dots, f_m\}$ in R_n: in the first case we have to reduce polynomials by the field equations, which is done at the same time as the multiplication when working in R_n. From now on, we consider only polynomials in R_n.

We recall some basic properties of Gröbner bases for equations over $\mathbb{F}_2$ with solutions in $\mathbb{F}_2$:

Proposition 1 *Let* $I \neq \{0\}$ *be an ideal in* R_n*, and* $<$ *an admissible monomial order, then*

- *there exists a Gröbner basis* G *of* I*, and* $\langle G \rangle = I$*,*
- *there exists a unique reduced Gröbner basis* G *of* I*, and* $G = \{1\}$ *iff* $V_{\overline{\mathbb{K}}}(I) = \emptyset$*.*
- *If* I *has only one solution in* $\mathbb{F}_2$*, say* $(a_1, \dots, a_n)$*, then* $G = \{x_1 + a_1, \dots, x_n + a_n\}$*.*

A good complexity measure is the index of regularity of the Hilbert function associated to the ideal. The Hilbert function of an ideal $I = \langle f_1, \dots, f_m \rangle$ is the function on the nonnegative integers d defined by $HF_{m,n}(d) = \dim(R_n/I)_d$ as a vector space (where F_d stands for $\{f \in F : \deg(f) = d\}$), hence $HF_{m,n}(d) = \dim((R_n)_d) - \dim(I_d)$. For d sufficiently large, the Hilbert function cancels (see [CLO97] for instance, this is because I is zero-dimensional), and the first d for which the Hilbert function cancels is called the index of regularity of I. The Hilbert series of I is $\sum_{d \geq 0} HF_{m,n}(d) z^d$, and this series is a polynomial. It is clear that for $d > n$ we have $(R_n/I)_d = (0)$ (there exists no square free

monomial in n variables of degree greater than n), hence the index of regularity is less than or equal to $n+1$. We will see in the next section that the maximal degree of an element in the Gröbner basis of I is less than the index of regularity, and is related to the global (arithmetic) cost of the computation when using linear algebra. One goal of this paper is to improve the bound $n+1$ for "generic" sequences.

2.2 Gröbner bases and linear algebra

We first consider homogeneous polynomials $f_1, \ldots, f_m$ in R_n^h the polynomial ring modulo the homogeneous part of largest degree of the field equations, that is $R_n^h = \mathbb{F}_2[x_1, \ldots, x_n]/(x_1^2, \ldots, x_n^2)$. Let $<$ be a graded admissible monomial order.

The vector space I_d is generated by all the tf_j with $1 \leq j \leq m$ and t a monomial such that $\deg(tf_j) = d$. So we can construct the Macaulay matrix $\mathcal{M}_{d,m}^{\text{acaulay}}$ as follows [Mac02]: write down horizontally all the degree d (square free) monomials, ordered following the monomial order $<$ (the first one being the largest one). Hence each column of the matrix is indexed by a monomial of degree at most d. Multiply each f_j from 1 to m by any monomial t of degree $d - d_j$ (hence the product tf_j has degree d), and write the coefficients of tf_j under their corresponding monomials, thus giving a row of the matrix. The rows are labeled (the row tf_j is labeled (t, f_j)) and ordered: row (t, f_j) is before row (u, f_i) if either $j < i$, or $j = i$ and $t < u$.

$$\mathcal{M}_{d,m}^{\text{acaulay}} = \begin{array}{c} \\ (t, f_1) \\ (u, f_2) \\ \vdots \\ (v, f_m) \end{array} \begin{array}{c} \text{monomials of degree } d \\ \begin{pmatrix} & & \\ & & \\ & & \\ & & \end{pmatrix} \end{array}$$

For any row in the matrix, consider the monomial indexing the first nonzero column of this row. It is called the leading monomial of the row, and is the leading monomial of the corresponding polynomial.

Gaussian elimination applied on this matrix leads to a Gröbner basis computation [Laz83, Laz01]. Indeed, call $\tilde{\mathcal{M}}_{d,m}^{\text{acaulay}}$ the Gaussian eliminated form of $\mathcal{M}_{d,m}^{\text{acaulay}}$, such that the only elementary operation allowed for one row is the addition of a linear combination of the previous rows. The labels of the rows in $\tilde{\mathcal{M}}_{d,m}^{\text{acaulay}}$ are the same as in $\mathcal{M}_{d,m}^{\text{acaulay}}$. Now consider all the polynomials corresponding to a row whose leading term is not the same in $\mathcal{M}_{d,m}^{\text{acaulay}}$ and $\tilde{\mathcal{M}}_{d,m}^{\text{acaulay}}$, for all $d \leq d_{\max}$, then this set of polynomials is a Gröbner basis of $f_1, \ldots, f_m$ up to degree $d_{\max}$, and for $d_{\max}$ large enough ($d_{\max} \leq n$), this set is a Gröbner basis.

For a sequence of affine polynomials $f_1, \ldots, f_m \in R_n$, define the Macaulay matrix $\mathcal{M}_{d,m}^{\text{acaulay}}$ to be the Macaulay matrix of the sequence $f_1^h, \ldots, f_m^h \in R_n^h$, where f_i^h is the homogeneous part of largest degree of f_i.

From the complexity point of view, if $d(n)$ is the maximal degree occurring in the computation, and $N_{d(n)}$ is the size of the largest matrix $\mathcal{M}_{d(n),m}$, then the whole complexity is dominated by the cost of the linear algebra on $\mathcal{M}_{d(n),m}$, which is $N_{d(n)}^{\omega}$ where ω is the coefficient of linear algebra, the best known bound being $\omega = 2.376$ [CW90]. Note that the matrix is very sparse: if the f_i's are of degree 2, then there are at most $\frac{n(n-1)}{2}$ non zero coefficients in each row, hence considering $\omega = 2$ is realistic. Considering the Macaulay matrix, many rows are zero in the reduced matrix $\tilde{\mathcal{M}}_{d(n),m}^{\text{acaulay}}$, and the number of rows is much larger than the number of columns and than the rank of the matrix. We will see in Section 4 that, using the criteria of the F5 Gröbner basis Algorithm, we

can extract from the Macaulay matrix a full rank matrix under some conditions on $f_1, \ldots, f_m$. In the next section, we define semi-regular sequences, for which the matrices in the F5 Algorithm will have full rank. The definition of semi-regular sequence is totally independent of any Gröbner basis algorithm and depends only on the ideal itself.

3 Definition of Semi-regular sequences

Let $f_1, \ldots, f_m$ be a sequence of m equations with degrees $d_1, \ldots, d_m$ over a field $\mathbb{K}$ in n variables $x_1, \ldots, x_n$, and $I = \langle f_1, \ldots, f_m \rangle$ the generated ideal. If $\mathbb{K} = \mathbb{Q}$, there already exists a notion of regular sequence, for which the complexity of the Gröbner basis computation is well known. Regular sequences can be defined by the set of relations between generators. They always verify the trivial relations $f_i f_j = f_j f_i$, as well as all the relations generated by these ones. Regular sequences are sequences which verify no other relations than the trivial ones. They can also be characterized by the dimension of the associated variety: a sequence $f_1, \ldots, f_m$ is regular if and only if $V(f_1, \ldots, f_m)$ has dimension $n - m$.

For overdetermined systems, the preceding definition does not work any more. We define semi-regular sequences and degree of regularity, that extend the notions of regular sequences and index of regularity to overdetermined systems. The degree of regularity is the first degree d for which the set $\{LT(f) : f \in I_d\}$ is exactly the set of monomials of degree d. Semi-regular sequences are sequences verifying no other relations than the trivial ones *up to degree* D_{reg}.

For systems with coefficients in $\mathbb{F}_2$ and solutions in $\mathbb{F}_2$, every polynomial $f \in I$ verifies $f^2 = f$, hence there is no semi-regular sequences with the previous definition. We extends the definition of semi-regular sequences by taking into account the trivial relations $f_i^2 = f_i$, hence semi-regular sequences over $\mathbb{F}_2$ are sequences verifying no other relations than the trivial ones $f_i f_j = f_j f_i$ and $f_i f_i = f_i$ up to degree D_{reg}. We give below the mathematical definition of semi-regular sequences over $\mathbb{F}_2$, beginning with the homogeneous case:

Definition 3 (Homogeneous $\mathbb{F}_2$-degree of regularity) *Let $f_1, \ldots, f_m$ be a sequence of m homogeneous polynomials in R_n^h, and $I = \langle f_1, \ldots, f_m \rangle$. We define the degree of regularity of I to be the minimal degree d such that $\{LT(f) : f \in I_d\}$ is exactly the set of monomials of degree d in R_n^h (or R_n), and denote it by D_{reg} (I). We have $D_{reg} \leq n + 1$.*

Remark: D_{reg} is exactly the index of regularity.

Definition 4 (Homogeneous $\mathbb{F}_2$-semi-regular sequences) *The sequence of homogeneous polynomials $f_1, \ldots, f_m \in R_n^h$ is a semi-regular sequence over $\mathbb{F}_2$ if*

- $I = \langle f_1, \ldots, f_m \rangle \neq R_n^h$
- *For $i = 1, \ldots, m$, if $g_i f_i = 0$ in $R_n^h/(f_1, \ldots, f_{i-1})$ and $\deg(g_i f_i) < D_{reg}$ (I) then $g_i = 0$ in $R_n^h/(f_1, \ldots, f_{i-1}, f_i)$*

An affine sequence of polynomials is said to be semi-regular over $\mathbb{F}_2$ if its homogeneous part is:

Definition 5 (Affine degree of regularity, semi-regular sequences over $\mathbb{F}_2$) *Let $f_1, \ldots, f_m$ be an affine sequence of polynomials in R_n and $I = \langle f_1, \ldots, f_m \rangle$. For each $1 \leq i \leq m$ let f_i^h be the homogeneous part of f_i of largest degree. The affine sequence $f_1, \ldots, f_m \in R_n$ is said to be semi-regular over $\mathbb{F}_2$ if the homogeneous sequence $f_1^h, \ldots, f_m^h \in R_n^h$ is semi-regular over $\mathbb{F}_2$. We define the degree of regularity of I to be the degree of regularity of I^h: D_{reg} (I)=D_{reg} (I^h).*

We conjecture that most sequences are semi-regular sequences:

Conjecture 2 *For any $(n, m, d_1, \ldots, d_m)$ the proportion of semi-regular sequences over $\mathbb{F}_2$ in the set $E(n, m, d_1, \ldots, d_m)$ of algebraic systems of m equations of degrees $d_1, \ldots, d_m$ in n variables tends to 1 as n tends to ∞.*

This conjecture implies that a "generic" sequence of polynomials is semi-regular. We did many computer experiments with random sequences, and we always got semi-regular sequences (with n big enough).

4 Analysis of Algorithm F5 and computation of $D_{\mathbf{reg}}$

The goal of this section is to analyze the behavior of Algorithm F5 [Fau02] when the input is a semi-regular sequence. We consider F5 in a matrix fashion: the algorithm matrix-F5 constructs matrices incrementally in the degree and the number of polynomials, and uses linear algebra. As stated in Section 2.2, the global cost is dominated by the cost of the linear algebra on the largest matrix, which is the matrix in degree D_{reg}.

After a short description of matrix-F5 Algorithm, we show that for semi-regular sequences all matrices are full rank and we give the exact value of this rank. This is closely related to the Hilbert series of semi-regular sequences as well as their degree of regularity.

4.1 Short description of matrix-F5 Algorithm

Let $f_1, \ldots, f_m$ be a sequence of m polynomials in the polynomial ring modulo the field equations, $R_n = \mathbb{F}_2[x_1, \ldots, x_n]/(x_1^2 + x_1, \ldots, x_n^2 + x_n)$, and $<$ a graded admissible monomial order. The rows of the Macaulay matrix $\mathcal{M}_{d,m}^{\mathrm{acaulay}}$ generate I_d as a vector space, but surely it has not full rank: we have at least the relations $f_i f_j = f_j f_i$ and $f_i f_i = f_i$. The idea of [Fau02] is to construct a sub-matrix[1] $\mathcal{M}_{d,m}$ of this matrix, incrementally in d, which will have full rank for semi-regular sequences.

The following proposition explicits the criterion used to construct $\mathcal{M}_{d,m}$ from $\mathcal{M}_{d,m}^{\mathrm{acaulay}}$ by removing all rows that will reduce to zero because of the trivial relations $f_i f_j = f_j f_i$ or $f_i f_i = f_i$:

Proposition 3 (Frobenius criterion) *if in the matrix $\tilde{\mathcal{M}}_{d-d_m,m}$ a row (t, f_m) has a leading term t' then the row (t', f_m) is redundant in the matrix $\mathcal{M}_{d,m}^{acaulay}$.*

Suppose we have constructed matrices $\mathcal{M}_{d',m}$ for $d' < d$. Then the matrix $\mathcal{M}_{d,m}$ consists of all the rows labeled (t, f_i) with $1 \leq i \leq m$ and $\deg(t) = d - d_i$ except for the t's that correspond to a (non-zero) leading term in $\tilde{\mathcal{M}}_{d-d_m,m}$. The number of total possibilities for t is the number of monomials of degree $d - d_i$, that is $\binom{n}{d-d_i}$, to which we remove as many rows as non-zero ones in $\tilde{\mathcal{M}}_{d-d_m,m}$. Hence, the number of rows in $\mathcal{M}_{d,m}$ is equal to $\binom{n}{d-d_i}$ minus the rank of $\tilde{\mathcal{M}}_{d-d_m,m}$.

Define $h_{d,m}(n)$ to be the number of columns in $\mathcal{M}_{d,m}$ minus the number of rows. As long as the matrices have full rank, and that is the case if $d < D_{\mathrm{reg}}$ for a semi-regular sequence, then $h_{d,m}(n)$ is the Hilbert function of the homogeneous part of highest degree of the sequence, and the previous reasoning leads to the recurrence formula:

[1] A matrix A is a sub-matrix of a matrix B if, for every row in A with label (u, f_i), there exists a row in B with the same label.

Lemma 4 *For a sequence $f_1, \ldots, f_m$, as long as there is no reduction to zero in matrix $\mathcal{M}_{d,m}$, then the Hilbert function $HF_{m,n}(d)$ satisfies the recurrence formula:*

$$HF_{m,n}(d) = HF_{m-1,n}(d) - HF_{m,n}(d - d_m)$$

with initial conditions $HF_{m,n}(d) = \binom{n}{d}$ if $m \leq 0$ or $d < \min(d_k : k \leq m)$.

Proposition 5 *If the sequence $f_1, \ldots, f_m$ is semi-regular, then all matrices $\mathcal{M}_{d,m}$ in Algorithm matrix-F5 with $d \leq D_{reg}$ are full rank.*

Proof Suppose a row (t', f_m) is zero in the matrix $\tilde{\mathcal{M}}_{d,m}$ with $d < D_{\mathrm{reg}}$. Then there exists a polynomial g with leading term t' such that $gf_m = 0$ in $R_n/(f_1, \ldots, f_{m-1})$ (the only allowed operations on the rows of the matrix are a linear combination of the previous rows). From the definition of semi-regular sequence, this implies that $g = 0$ in $R_n/(f_1, \ldots, f_m)$, hence t' is the leading term of a row in $\tilde{\mathcal{M}}_{d-d_m,m}$ (the rows of this matrix generates I_{d-d_m} as a vector space), and from the criteria this row has been removed from the matrix $\mathcal{M}_{d,m}$. For $d = D_{\mathrm{reg}}$, by definition of D_{reg} the rank of $\mathcal{M}_{d,m}$ is exactly the number of columns, hence the matrix has full rank. □

This means that, for semi-regular sequences, the number of rows of all matrices in degree at most D_{reg} is known, and is given by a recurrence formula.

4.2 Generating series

Generating series are very powerful objects to study sequences defined by a recurrence formula. Here the generating series associated to the recurrence formula of Lemma 4 is very simple, and as long as there is no reduction to zero, then $HF_{m,n}(d)$= # columns - # rows is given by the coefficient of degree d of these series, and is very easy to compute.

Proposition 6 *Let $h_{d,m}(n)$ be a sequence satisfying the recurrence formula of Lemma 4, that is to say $h_{d,m}(n) = h_{d,m-1}(n) - h_{d-d_m,m}(n)$ with initial conditions $h_{d,m}(n) = \binom{n}{d}$ if $m \leq 0$ or $d < \min(d_k : k \leq m)$. Then the generating series of $h_{d,m}(n)$ is*

$$S_{m,n}(z) = \sum_{d \geqslant 0} h_{d,m}(n) z^d = (1+z)^n \Big/ \prod_{k=1}^{m} \left(1 + z^{d_k}\right). \tag{1}$$

Corollary 7 *The Hilbert series of a semi-regular sequence of m polynomials with degrees $d_1, \ldots, d_m$ over $\mathbb{F}_2$ is*

$$\left[(1+z)^n \Big/ \prod_{i=1}^{m} (1 + z^{d_i})\right] \tag{2}$$

where $\left[\sum_{i \geq 0} a_i z^i\right] = \sum_{i \geq 0} b_i z^i$ with $b_i = a_i$ if $a_j > 0 \; \forall 0 \leq j \leq i$ and $b_i = 0$ otherwise. Conversely, any sequence of m polynomials with degrees $d_1, \ldots, d_m$ and Hilbert series (2) is semi-regular.

4.3 Explicit computation of D_{reg} for semi-regular sequences

The following lemma enables us to compute the degree of regularity D_{reg} from the generating series:

Lemma 8 *Let $f_1, \ldots, f_m$ be a semi-regular over $\mathbb{F}_2$, and denote by*

$$S_{m,n}(y) = \sum_{d \geqslant 0} h_{d,m}(n) z^d$$

its generating series. Then we have $\forall d \leq D_{reg}, \quad HF_{m,n}(d) = h_{d,m}(n)$ and D_{reg} is characterized by

$$\forall d < D_{reg}, \quad h_{d,m}(n) > 0 \quad \text{and} \quad h_{D_{reg},m}(n) \leq 0$$

Proof As the sequence is semi-regular, for $d <D_{\rm reg}$ there is no reduction to zero so that we have more columns than rows and $HF_{m,n}(d) = h_{d,m}(n) > 0$ (it can not be equal to zero by definition of $D_{\rm reg}$). For $d =D_{\rm reg}$, all monomials of degree $D_{\rm reg}$ are reached, so that the rank of the matrix is exactly the number of columns, and less than or equal to the number of rows, hence $h_{D_{\rm reg},m}(n) = \#\text{columns} - \#\text{ rows } \leq \#\text{columns} - \text{ rank } = 0$. □

Thus, we get an easy way to compute $D_{\rm reg}$ using the generating series: we just have to evaluate the first coefficients of the series until the first negative one.

Let us consider for example n quadratic equations over $\mathbb{F}_2$. The generating series is

$$\begin{aligned} S_{n,n}(z) = \left(\frac{1+z}{1+z^2}\right)^n &= 1 + nz + \frac{1}{2} n\left(n-3\right) z^2 + \frac{1}{6} n\left(n^2 - 9\,n + 2\right) z^3 \\ &+ \frac{1}{24} n\left(n^3 - 18\,n^2 + 35\,n + 6\right) z^4 \\ &+ \frac{1}{120} n\left(-30\,n + 24 - 30\,n^3 + 155\,n^2 + n^4\right) z^5 + O\left(z^6\right) \end{aligned}$$

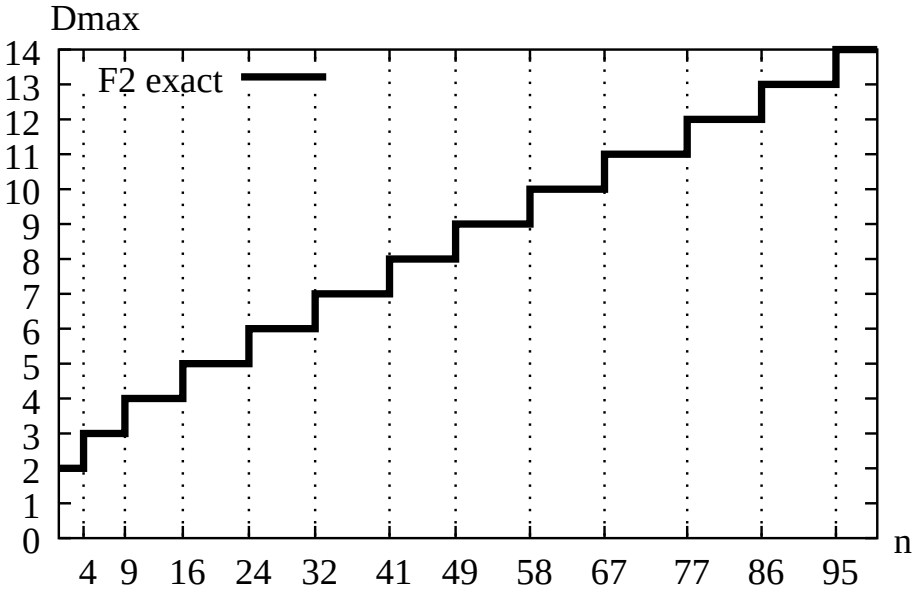

Figure 1: $D_{\rm reg}$ for $m = n$ quadratic polynomials over $\mathbb{F}_2$

The coefficient of z is n and is always positive, so that we always have $D_{\rm reg} \geq 1$. The coefficient of z^2 is ≤ 0 for $n \leq 3$, and > 0 for $n > 3$, so that we get $D_{\rm reg} = 2$ for $n \leq 3$ and $D_{\rm reg} \geq 3$ for $n > 3$.

The coefficient of z^3 has 2 roots $\frac{9\pm\sqrt{73}}{2} \simeq 8.8, 0.2$, so for $3 < n \leq 8$ it is negative and $D_{\text{reg}} = 3$, and positive for $n \geq 9$, and $D_{\text{reg}} \geq 4$. The two largest root of the coefficient in z^4 are about 15.75 and 2.4, for $9 \leq n \leq 15$ we get $D_{\text{reg}} = 4$ and for $n \geq 16$ we have $D_{\text{reg}} \geq 5$. We can successively compute the largest root of the coefficient of z^d in $S_{n,n}(z)$ to plot the stairs of Figure 1 page 10.

The question that arises then is: what is the slope of this graph? the answer is given by our computation of the asymptotic expansion of D_{reg} in the next section.

5 Asymptotic analysis of D_{reg}

We use asymptotic methods to determine the asymptotic behavior of D_{reg} in terms of n. We refer to [dB81, CFU57, Olv97] for background knowledge on the saddle points and the coalescent saddle points methods.

5.1 Methods

We deduced in Section 4.3 that, for semi-regular sequences, D_{reg} is the smallest d for which $h_{d,m}(n) = [z^d]S_{m,n}(z)$ is nonnegative. This leads us to consider $h_{d,m}(n)$, which is a polynomial in n, and to study its largest root in terms of d (then d in terms of n is exactly D_{reg}). For that purpose, we write $h_{d,m}(n)$ as an integral in the complex plane, and with a change of variable $\lambda = \frac{n}{d}$ the integral has the shape:

$$I_d(\lambda) = \oint g(z)e^{df(z,\lambda)}dz. \tag{3}$$

We are interested in the largest value of λ for which this integral cancels. We choose to compute the first term of the asymptotic expansion of $I_d(\lambda)$ with λ a parameter. We use the saddle point method: the saddle points are the roots of $f'(z,\lambda) = \frac{df(z,\lambda)}{dz}$, and the value of the integral is concentrated around the saddle points. The value of I is equal to the sum of the contributions of the saddle points: for a simple saddle point ζ, the contribution is $\frac{g(\zeta)}{\sqrt{2\pi d|f''(\zeta,\lambda)|}}e^{df(\zeta,\lambda)}(1+O(d^{-1}))$ whereas for a saddle point with multiplicity 2, it is $\left(\frac{2}{27d|f'''(\zeta,\lambda)|}\right)^{1/3} g(\zeta)\Gamma\left(\frac{1}{3}\right)e^{df(\zeta,\lambda)}(1+O(d^{-1}))$.

If the integral cancels for a value of λ, in particular the first term of the asymptotic expansion of $I_d(\lambda)$ is equal to zero. Hence λ has to be found among either the values that annihilate the coefficient in $\frac{1}{\sqrt{d}}$, or the values for which the main saddle points have multiplicities at least 2, which corresponds to the values of λ that annihilate the discriminant of the saddle points equation $f'(z,\lambda) = 0$. This gives the first term λ_0 of the asymptotic expansion of λ in terms of d, or equivalently of D_{reg} in terms of n. The choice of the variable λ rather than n is justified by the fact that we find λ_0 finite (and not zero).

To get the next term of λ in terms of d, we need to compute the asymptotic expansion of $I_d(\lambda)$ for λ in a neighborhood of λ_0, uniformly in λ. This cannot be done by the saddle point method if for λ_0 we have a double saddle point and for $\lambda \neq \lambda_0$ we have two simple saddle points. In that case, we use the coalescent saddle point method by Chester, Friedman and Ursell [CFU57], which gives the full asymptotic expansion of $I_d(\lambda)$ in terms of d and the Airy function Ai. As before, by successively computing and annihilating the first term, we get the whole asymptotic expansion of λ in terms of n.

5.2 An explicit example

Let us apply the previous method in case of n quadratic equations over the field $\mathbb{F}_2$. The generating series is $S_{n,n}(z) = \left(\frac{1+z}{1+z^2}\right)^n$. This gives $I_d(\lambda) = \oint \left(\frac{(1+z)}{1+z^2}\right)^n \frac{dz}{z^d} = \oint e^{df(z)}dz$ with $f(z) = \lambda \log(1+z) - \lambda \log(1+z^2) - \log(z)$. The derivative of f is $\frac{df}{dz} = -\frac{(\lambda+1)z^3+(2\lambda+1)z^2+(1-\lambda)z+1}{(z+1)(1+z^2)z}$ and there are three saddle points. One is real negative and its contribution to the integral is negligible compared to the contribution of the two other saddle points, denoted by $z_0^{\pm}$. The discriminant of the numerator of $\frac{df}{dz}$ is $\Delta = 8\lambda^4 - 80\lambda^3 - 96\lambda^2 - 32\lambda - 16$, it has one real positive root $\lambda_0 = \frac{3\sqrt{3}+5+\sqrt{6(12+7\sqrt{3})}}{2} \simeq 11.11$. For $\lambda > \lambda_0$, both saddle points are real, whereas for $0 < \lambda < \lambda_0$ they are complex conjugate and for $\lambda = \lambda_0$ they coalesce into a single saddle point of order 2.

As long as $\lambda \neq \lambda_0$, the two saddle points are distinct, and the asymptotic expansion of $I_d(\lambda)$ can be computed as before as the sum of the contribution of each saddle point, to give:

$$I_d(\lambda) \sim \left(\sqrt{\frac{2\pi}{f_{zz}(z_0^+)}}e^{df(z_0^+)} + \sqrt{\frac{2\pi}{f_{zz}(z_0^-)}}e^{df(z_0^-)}\right)\frac{1}{\sqrt{d}} \sim \sqrt{\frac{2\pi}{df_{zz}(z_0^+)}}e^{df(z_0^+)} \text{ if } \lambda > \lambda_0$$
$$\sim 2\Re\left(\sqrt{\frac{2\pi}{df_{zz}(z_0^+)}}e^{df(z_0^+)}\right) \text{ if } 0 < \lambda < \lambda_0$$

which is never zero. We deduce that the first term of the asymptotic expansion of $\lambda = \frac{n}{d}$ is $\lambda = \lambda_0 + o(1)$, hence the first term of the asymptotic expansion of D_{reg} in terms of n is $D_{\text{reg}} = \frac{n}{\lambda_0} + o(n)$.

To get the next term of the asymptotic expansion of λ in the neighborhood of λ_0, it is desirable to get an asymptotic expansion of $I_d(\lambda)$ which remains uniformly valid in a neighborhood of λ_0. For that purpose, we follow the method of coalescent saddle points by Chester, Friedman, and Ursell [CFU57], which is exposed in details in [dB81, Olv97] for instance. In order to estimate asymptotically the integral $I_d(\lambda)$, we introduce the *cubic* change of variable

$$f(z,\lambda) = \frac{u^3}{3} - \zeta(\lambda)u + \eta(\lambda) = \frac{u^3}{3} - \zeta u + \eta \tag{4}$$

where

$$\zeta^{\frac{3}{2}} = \frac{3}{4}(f(z_0^-,\lambda) - f(z_0^+,\lambda)) \text{ and } \eta = \frac{1}{2}(f(z_0^-,\lambda) + f(z_0^+,\lambda))$$

It is proved in [CFU57] that, writing $\frac{dz}{du} = \sum_{i\geq 0}(u^2-\zeta)^i(c_i + b_i u)$, the resulting integral can be expressed in terms of the Airy function:

$$I_d(\lambda) \sim e^{d\eta(\lambda)} 2I\pi \left[\frac{\mathrm{Ai}(d^{\frac{2}{3}}\zeta)}{d^{\frac{1}{3}}}\sum_{m\geq 0}\frac{B_m}{d^m} + \frac{\mathrm{Ai}'(d^{\frac{2}{3}}\zeta)}{d^{\frac{2}{3}}}\sum_{m\geq 0}\frac{C_m}{d^m}\right] \tag{5}$$

where Ai is the Airy function, and the B_m and C_m coefficients can be expressed in terms of the coefficients b_m and c_m. The dominant term in (5) is a multiple of $\mathrm{Ai}(d^{\frac{2}{3}}\zeta)$ which cancels when $d^{\frac{2}{3}}\zeta = a_1$ the first root of the Airy function. Putting $\lambda = \lambda_0 - d_\lambda^2$, we get $\zeta = \zeta_2 d_\lambda^2(1+o(1))$ and equation $d^{\frac{2}{3}}\zeta = a_1$ gives $d_\lambda^2 = \frac{a_1}{\zeta_2}\frac{1}{d^{\frac{2}{3}}}$, and hence $\lambda = \lambda_0 - \frac{a_1}{\zeta_2}\frac{1}{d^{\frac{2}{3}}} + o(\frac{1}{d^{\frac{2}{3}}})$.

By successively determining the dominant term in (5) and annihilating it, we obtain as many terms as needed for the asymptotic expansion of λ in terms of d. Here we give the first four terms:

$$\begin{aligned}\lambda &= \lambda_0 + \frac{a_1 \sqrt[3]{\frac{5}{6}\,{\lambda_0}^3 + \frac{1}{2}\,{\lambda_0}^2 + \frac{1}{6}}}{d^{2/3}} + \frac{\sqrt[3]{\frac{265}{648}\,{\lambda_0}^3 - \frac{85}{36}\,{\lambda_0}^2 - \frac{16}{27}\,\lambda_0 - \frac{38}{81}}}{d} \\ &\quad + {a_1}^2 \sqrt[3]{\frac{1531}{273375}\,{\lambda_0}^3 + \frac{9917}{729000}\,{\lambda_0}^2 + \frac{3163}{729000}\,\lambda_0 + \frac{1034}{273375}}\,\frac{1}{d^{\frac{4}{3}}} + O\left(\frac{1}{d^{\frac{5}{3}}}\right) \\ &= \lambda_0 + \frac{\lambda_1}{d^{2/3}} + \frac{\lambda_2}{d} + \frac{\lambda_3}{d^{4/3}} + O\left(\frac{1}{d^{5/3}}\right) \\ &= 11.114 - \frac{24.886}{d^{\frac{2}{3}}} + \frac{6.4043}{d} + \frac{11.545}{d^{\frac{4}{3}}} + O\left(\frac{1}{d^{\frac{5}{3}}}\right).\end{aligned}$$

Reverting this expansion, we get:

Proposition 9 *The first terms of the asymptotic expansion of D_{reg} for n quadratic equations in n variables is:*

$$\begin{aligned}D_{reg} &= \frac{n}{\lambda_0} - \frac{\lambda_1}{{\lambda_0}^{4/3}} n^{1/3} - \left(\frac{\lambda_2}{\lambda_0} + 1\right) + \left(\frac{{\lambda_1}^2}{3{\lambda_0}^{5/3}} - \frac{\lambda_3}{{\lambda_0}^{2/3}}\right)\frac{1}{n^{1/3}} + O\left(\frac{1}{n^{2/3}}\right) \\ &= 0.0900\,n + 1.00\,n^{\frac{1}{3}} - 1.58 + \frac{1.41}{n^{\frac{1}{3}}} + O\left(\frac{1}{n^{2/3}}\right).\end{aligned}$$

The asymptotic expansion is valid even for small n. Figure 2 page 14 compares the exact values of D_{reg}, computed from the generating series, and the asymptotic formulas with two or three terms. We can see that the asymptotic formula for D_{reg} is also a very good approximation of D_{reg} for small values of n. Of course, the fourth term in $n^{-\frac{1}{3}}$ is not small for small n, and the asymptotic expansion with four terms is not close to the exact value of D_{reg}.

Comparison with $2n$ equations over $\mathbb{Q}$ A system of n quadratic equations over $\mathbb{F}_2$ consists in fact of $2n$ quadratic equations, as we add the field equations $x_1^2 + x_1, \ldots, x_n^2 + x_n$. It is then interesting to compare the behavior of n quadratic equations over $\mathbb{F}_2$ with this of $2n$ quadratic equations over $\mathbb{Q}$, the difference being that, with the field equations, every polynomial of the ideal is solution of the relation $f^2 = f$ (action of the Frobenius morphism), which is not true for semi-regular sequences over $\mathbb{Q}$.

Over $\mathbb{Q}$, for $2n$ quadratic equations we get

$$\begin{aligned}\lambda &= \lambda_0 + a_1 \sqrt[3]{-\frac{99}{2} + \frac{577}{4}\,\lambda_0}\,\frac{1}{d^{\frac{2}{3}}} + \sqrt[3]{-\frac{309}{64} + \frac{1855}{128}\,\lambda_0}\,\frac{1}{d} \\ &\quad + a_1^2 \sqrt[3]{-\frac{899557}{2304000} + \frac{15909029}{13824000}\,\lambda_0}\,\frac{1}{d^{\frac{4}{3}}} + O\left(\frac{1}{d^{\frac{5}{3}}}\right) \\ &= 11.657 - \frac{27.528}{d^{\frac{2}{3}}} + \frac{5.4749}{d} + \frac{12.862}{d^{\frac{4}{3}}} + O\left(\frac{1}{d^{\frac{5}{3}}}\right) \\ D_{\mathrm{reg}} &= 0.0858\,n + 1.04\,n^{\frac{1}{3}} - 1.47 + \frac{1.71}{n^{\frac{1}{3}}} + O\left(\frac{1}{n^{\frac{2}{3}}}\right)\end{aligned}$$

The degree of regularity D_{reg} is a little smaller over $\mathbb{Q}$ than over $\mathbb{F}_2$ (over $\mathbb{F}_2$ we have $D_{\mathrm{reg}} \sim n/11.11$ whereas over $\mathbb{Q}$ we have $D_{\mathrm{reg}} \sim n/11.657$). This is due to the reductions to zero coming from the

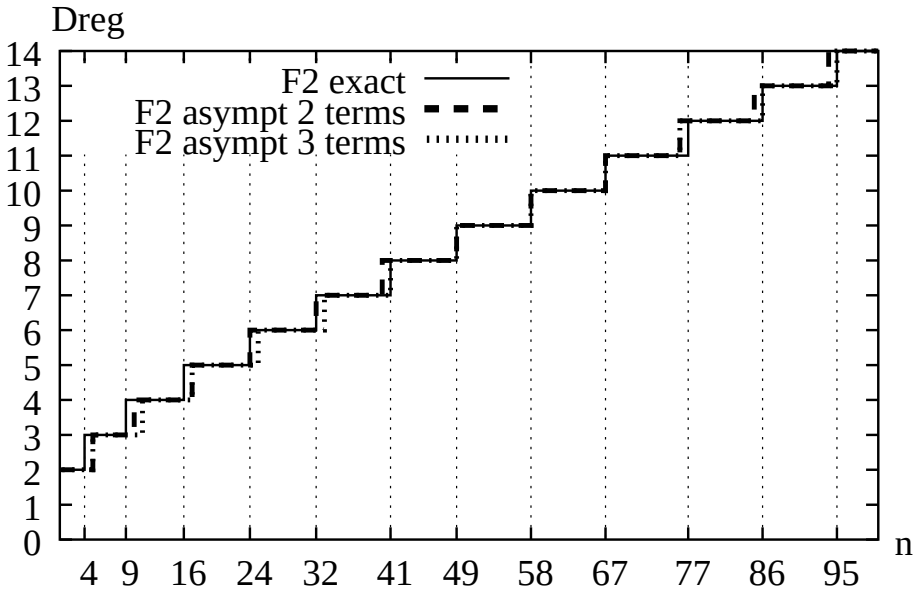

Figure 2: D_{reg} for $m = n$ quadratic polynomials over $\mathbb{F}_2$, exact and asymptotic values

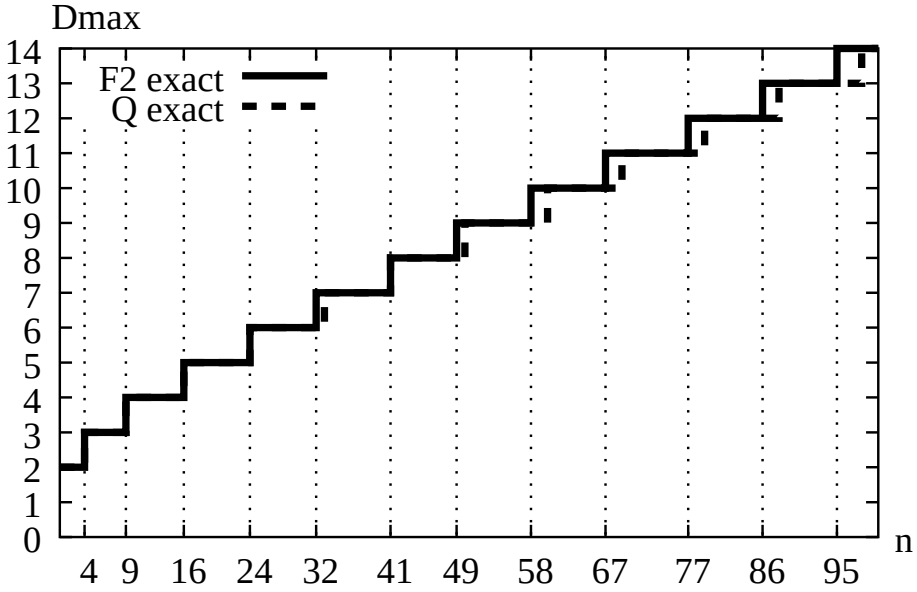

Figure 3: D_{reg} for $m = n$ quadratic polynomials over $\mathbb{F}_2$ and $m = 2n$ quadratic polynomials over $\mathbb{Q}$

Frobenius morphism. Figure 3 page 14 compares D_{reg} in both cases, and it shows that for $n < 100$ the difference is small (D_{reg} differs from at most 1, and $n = 295$ is the smallest n for which the difference is 2, we have $D_{\mathrm{reg}} = 32$ over $\mathbb{Q}$ and $D_{\mathrm{reg}} = 34$ over $\mathbb{F}_2$).

The field equations behave approximatively as other equations from the complexity point of view.

The degree of regularity for different D. We give below the asymptotic value of $D_{\rm reg}$ for n equations of degree D over $\mathbb{F}_2$, for different values of D:

Proposition 10 *For $m = n$ equations of degree D over $\mathbb{F}_2$, with different values of D we get*

D	*asymptotic*
2	$D_{reg} = .09n + 1.00n^{\frac{1}{3}} - 1.58 + O(\frac{1}{n^{\frac{1}{3}}})$
3	$D_{reg} = .15n + 1.35n^{\frac{1}{3}} - 1.42 + O(\frac{1}{n^{\frac{1}{3}}})$
4	$D_{reg} = .20n + 1.60n^{\frac{1}{3}} - 1.27 + O(\frac{1}{n^{\frac{1}{3}}})$
5	$D_{reg} = .24n + 1.79n^{\frac{1}{3}} - 1.11 + O(\frac{1}{n^{\frac{1}{3}}})$
6	$D_{reg} = .26n + 1.95n^{\frac{1}{3}} - 0.94 + O(\frac{1}{n^{\frac{1}{3}}})$
7	$D_{reg} = .28n + 2.09n^{\frac{1}{3}} - 0.78 + O(\frac{1}{n^{\frac{1}{3}}})$

As for $D = 2$, all coefficients are explicit and can be expressed as algebraic functions of the first root a_1 of the Airy function and the algebraic number λ_0.

6 Complexity classification for quadratic equations over $\mathbb{F}_2$

Consider a sequence of m quadratic equations in n variables over $\mathbb{F}_2$, with generating series $(1+z)^n/(1+z^2)^m$. The saddle point equation is $f_z(z) = \frac{n}{1+z} - \frac{2mz}{1+z^2} - \frac{d}{z}$; its numerator is $(n-d-2m)z^3 - (d+2m)z^2 + (n-d)z - d$. There are three saddle points: one does not contribute to the asymptotic expansion of the integral (its contribution is negligible compared to the contribution of the two other ones), and the two saddle points coalesce as the discriminant of this equation equals zero, that is as $-4\,d^4+(8\,n-16\,m)\,d^3+\left(-20\,m^2+4\,m\,n-8\,n^2\right)d^2+\left(-8\,m^3+4\,n^3+4\,m\,n^2-20\,m^2n\right)d+m^2n^2+2\,m\,n^3-n^4=0$. This equation has exactly one real positive root in d if $m > \frac{n}{4}$, and no real positive root in all other cases. The solution is

$$d = \frac{n}{2} - m + m\sqrt{-\left(\frac{n}{m}\right)^2 - 10\frac{n}{m} + 2 + 2\sqrt{8\left(\frac{n}{m}\right)^3 + 12\left(\frac{n}{m}\right)^2 + 6\frac{n}{m} + 1}} \qquad (6)$$

In this Section we use without proof the fact that the coalescent saddle point method applies here for $m > \frac{n}{4}$: we assume that the first term of the asymptotic expansion of $D_{\rm reg}$ is the value of the double saddle point, i.e. $D_{\rm reg}$ is asymptotically equivalent to the value (6). We determine the behavior of $D_{\rm reg}$ as m ranges from $n/4$ to n^2, as well as the asymptotic expansion of $\text{Size} = \binom{n}{D_{\rm reg}}$ the size of the last matrix in degree $D_{\rm reg}$. Then the global cost of the Gröbner basis computation is of order Size^ω where ω is the coefficient of linear algebra.

Case $\frac{m}{n}$ tends to a constant. We have $m = n(N + o(1))$ with $N \geq 1/4$. In that case,

$$\begin{aligned}\frac{D_{\rm reg}}{n} &= \frac{1}{2} - N + \frac{1}{2}\sqrt{2N^2 - 10N - 1 + 2(N+2)\sqrt{N(N+2)}} + o(1)\\ &= D_0 + o(1) \text{ with } 0 < D_0 \leq 1/4\end{aligned}$$

so that the size of the last matrix is:

$$\log(\text{Size}) = n[-(1-D_0)\log(1-D_0) - D_0\log(D_0)](1+o(1)) = nD_1(1+o(1))$$

and the global cost of the Gröbner basis computation is $\left(2^{D_1/\log(2)}\right)^{n\omega}$ which is *exponential*. Note that $0 < D_1 \leq 2\log(2) - 3/4\log(3) < \log(2)$.

Case $\frac{m}{n^2}$ tends to a constant. Put $m = Nn^2(1+o(1))$, then we get

$$D_{\text{reg}} = \frac{n^2}{8m}(1+o(1)) = \frac{1}{8N} + o(1)$$

Hence the global cost of the Gröbner basis computation is *polynomial*: Size $\sim \max(n^2, n^{1/(8N)})$ (we have at least $D_{\text{reg}} \geq 2$).

Case $\frac{m}{n}$ tends to ∞ and $\frac{m}{n^2}$ tends to 0. As previously, we have $D_{\text{reg}} = \frac{n^2}{8m}(1+o(1)) = o(n)$. We use the fact that, if $A, B \to \infty$ and $A = o(B)$, then $\log\binom{B}{A} = A\log\left(\frac{B}{A}\right) + A + o(A)$ and apply that for $B = n$ and $A = D_{\text{reg}} = o(n)$:

$$\log(\text{Size}) \sim \frac{n}{8}\frac{\log(m/n)}{m/n} = o(n)$$

and the global cost is *sub-exponential*. For instance, if $m = n\log(n)$ equations, the global cost is $e^{\omega\frac{n}{8}\frac{\log(\log(n))}{\log(n)}}$.

Summary. For quadratic semi-regular sequences, the global cost of the Gröbner basis computation can be classified as follows:

$m(n)$ polynomials	degree of regularity Δ (n)	Size of the matrix	Global Complexity
$\sim Nn$ ($N \geq 1/4$)	$\sim D_0 n$ $0 < D_0 \leq 1/4$	$\sim \left(2^{D_1/\log(2)}\right)^n$ with $0 < D_1 \leq 2\log(2) - 3/4\log(3)$	exponential
$n \ll m \ll n^2$	$\sim \frac{n^2}{8m} = o(n)$	$\sim \frac{n}{8}\frac{\log(m/n)}{m/n} = o(n)$	sub-exponential
$\sim Nn^2$	$\sim 1/(8N)$	$\sim n^{1/(8N)}$	polynomial

Note that the behavior of the Gröbner basis computation for m quadratic equations in n variables is always *at most* polynomial in 2^n, hence *at most exponential*. Given a system of m quadratic equations, its complexity corresponds to the prediction from this table with a good probability. This table gives "a priori" upper bounds.

7 Some Applications

Equations over $\mathbb{F}_q$ with q large. For equations with coefficients in a field $\mathbb{F}_q$, with q large, the field equations have a degree too large to be useful. Hence we can forgot them and the complexity is the same as if the equations where with coefficients in $\mathbb{Q}$. It can be shown that the Hilbert series of a semi-regular sequence of m equations in n variables with degrees $d_1, \ldots, d_m$ without field equations is $\prod_{i=1}^{m}(1-z^{d_i})\Big/(1-z)^n$, and D_{reg} can be computed as before.

Application to the HFE problem. Sometimes the Gröbner basis computation is easier than the prediction, and this is the case for the HFE public key cryptosystem [Pat96].

The public key is an algebraic system of n quadratic equations in n variables over $\mathbb{F}_2$. The first challenge was proposed in [Pat96], and consisted of a system of 80 equations in 80 unknowns. We can build the table of D_{reg} for n quadratic equations in n variables, and the size of the largest matrices $\mathcal{M}_{D_{\text{reg}},n}$ is given by the generating series:

$$\sum_{d\geq 0}(\#\text{rows})z^d = (1+z)^n - \left(\frac{1+z}{1+z^2}\right)^n \tag{7}$$

We construct from that formula the following table:

n	15	16	23	24	76	77	...	80
D_{reg}	4	5	5	6	11	12		12
# rows in $\mathcal{M}_{D_{\text{reg}},n}$	1455	6784	34385	174824	$\sim 2^{42}$	$\sim 2^{45}$		$\sim 2^{46}$

From the table we get $D_{\text{reg}} = 12$ for the parameters of the first HFE challenge, and a semi-regular sequence of 80 equations in 80 unknowns would lead to solve a linear system of size 2^{46}. Even with sparse linear algebra, the global cost is greater than 2^{92}, which is unfeasible. It is shown in [FJ03] that the first HFE challenge was solvable, with the maximal degree occuring during the computation being only 4, and the size of the matrix in degree 4 being a 307126×1667009 matrix. Hence the HFE challenge was not a semi-regular sequence and was much easier than semi-regular sequences.

The prediction from the table is far away from the true maximal degree, but is still useful up to degree $d_{\max} < D_{\text{reg}}$ if there is no reduction to zero up to degree $d_{\max}$. We can define a notion of $d_{\max}-$semi-regular sequences, which are sequences that are semi-regular *up to degree* $d_{\max}$. With this definition, semi-regular sequences are $D_{\text{reg}}-$semi-regular sequences. Then the size of all matrices in F5 up to degree $d_{\max}$ are predictable, and given by the same formula as for semi-regular sequences, and we can predict the behavior of F5 for the first degrees. The maximal degree $d_{\max}$ being given (from experiments for instance), we can compute the size of the greatest matrix and estimate the global cost of the resolution of the system.

Estimation of the complexity for the algebraic systems derived in [BDC03] In [BDC03], the authors derived for many cryptosystems a set of algebraic equations verified by the input and output bits, the question being to estimate the complexity of the resolution of such systems. Applying our results, we find upper bounds for their attacks.

We get from [BDC03] the following table, giving the number of equations found among the input bits and the output bits for different cryptosystems:

	Khazad	Misty1	Kasumi	Camellia-128	Rijndael-128	Serpent-128
Variables	6464	3856	4264	3584	3296	16640
Linear eqs.	1664	2008	2264	1920	1696	8320
Quadratic eqs.	6000	1848	2000	4304	4600	9360

From that table, we deduce the following one, where the number m of equations takes into account the n field equations and the linear equations have been removed:

	Khazad	Misty1	Kasumi	Camellia-128	Rijndael-128	Serpent-128
n	4800	1848	2000	1664	1600	8320
m	10800	3696	4000	5968	6200	17680
D_{reg}	237	106	114	60	55	423
Size of $\mathcal{M}_{D_{\text{reg}},m}$	2^{1357}	2^{581}	2^{626}	2^{368}	2^{341}	2^{2407}
a priori Global Complexity($\omega = 2$)	2^{2713}	2^{1162}	2^{1252}	2^{737}	2^{682}	2^{4814}

Those systems are clearly unsolvable. It does not mean that an algebraic attack will never work, neither that this attack does not work. It just means that this attack will work only if those systems are much easier than semi-regular systems. The authors limited themselves to non-overdetermined systems, but we have shown that overdetermined systems are *in general* easier to solve than non-ones, of course if the additionnal equations are independant enough from the first ones.

References

[AK03] F. Armknecht and M. Krause. Algebraic Attacks on Combiners with Memory. In *Advances in Cryptology - CRYPTO 2003*, vol. 2729 of *Lecture Notes in Computer Science*, pp. 162 – 175. Springer-Verlag, Heidelberg, 2003.

[BDC03] A. Biryukov and C. De Cannière. Block Ciphers and Systems of Quadratic Equations. In *Proceedings of Fast Software Encryption*, pp. 291–306, 2003.

[Buc65] B. Buchberger. *Ein Algorithmus zum Auffinden der Basiselemente des Restklassenringes nach einem nulldimensionalen Polynomideal*. Ph.D. thesis, Innsbruck, 1965.

[Buc70] B. Buchberger. Ein algorithmisches Kriterium für die Lösbarkeit eines algebraischen Gleichungssystems. *Aequationes Math.*, vol. 4: pp. 374–383, 1970.

[CFU57] C. Chester, B. Friedman and F. Ursell. An extension of the method of steepest descents. *Proc. Camb. Philos. Soc.*, vol. 53: pp. 599–611, 1957.

[CLO97] D. Cox, J. Little and D. O'Shea. *Ideals, varieties, and algorithms*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, second ed., 1997. An introduction to computational algebraic geometry and commutative algebra.

[CM03] N. T. Courtois and W. Meier. Algebraic Attacks on Stream Ciphers with Linear Feedback. In *Advances in Cryptology – EUROCRPYT '03 (Warsaw, Poland, 2003)*, vol. 2656 of *Lecture Notes in Computer Science*, pp. 345 – 359. Springer-Verlag, Heidelberg, 2003.

[CW90] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. Symbolic Comput.*, vol. 9(3): pp. 251–280, 1990.

[dB81] N. G. de Bruijn. *Asymptotic methods in analysis*. Dover Publications Inc., New York, third ed., 1981.

[Fau02] J.-C. Faugère. A new efficient algorithm for computing Gröbner bases without reduction to zero (F_5). In T. Mora, editor, *ISSAC 2002*, pp. 75–83, 2002.

[FJ03] J.-C. Faugère and A. Joux. Algebraic Cryptanalysis of Hidden Field Equation (HFE) Cryptosystems Using Gröbner bases. In D. Boneh, editor, *Advances in Cryptology - CRYPTO 2003*, vol. 2729 of *LNCS*, pp. 44–60. Springer, 2003.

[FY80] A. S. Fraenkel and Y. Yesha. Complexity of solving algebraic equations. *Inform. Process. Lett.*, vol. 10(4-5): pp. 178–179, 1980.

[Kob98] N. Koblitz. *Algebraic aspects of cryptography*, vol. 3 of *Algorithms and Computation in Mathematics*. Springer-Verlag, Berlin, 1998. With an appendix by Alfred J. Menezes, Yi-Hong Wu and Robert J. Zuccherato.

[Laz83] D. Lazard. Gröbner bases, Gaussian elimination and resolution of systems of algebraic equations. In *Computer algebra (London, 1983)*, vol. 162 of *Lecture Notes in Comput. Sci.*, pp. 146–156. Springer, Berlin, 1983.

[Laz01] D. Lazard. Solving systems of algebraic equations. *ACM SIGSAM Bulletin*, vol. 35(3): pp. 11–37, 2001.

[Mac02] F. S. Macaulay. On some formula in elimination. In *London Mathematical Society*, no. 33 in 1, pp. 3–27, 1902.

[MI88] T. Matsumoto and H. Imai. Public quadratic polynomial-tuples for efficient signature-verification and message-encryption. In *Advances in cryptology – EUROCRYPT '88 (Davos, 1988)*, vol. 330 of *Lecture Notes in Comput. Sci.*, pp. 419–453. Springer, Berlin, 1988.

[MM82] E. W. Mayr and A. R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Advances in Mathematics*, vol. 46(3): pp. 305–329, 1982.

[Moh99] T. T. Moh. A Public Key System With Signature And Master Key Functions. *Communications in Algebra*, vol. 27(5): pp. 2207–2222, 1999.

[MR02] S. Murphy and M. J. B. Robshaw. Essential Algebraic Structure within the AES. In M. Yung, editor, *Advances in Cryptology – CRYPTO 2002*, vol. 2442 of *Lecture Notes in Computer Science*, pp. 1–16. Springer, Berlin, 2002.

[Olv97] F. W. J. Olver. *Asymptotics and special functions*. AKP Classics. A K Peters Ltd., Wellesley, MA, 1997. Reprint of the 1974 original [Academic Press, New York; MR **55** #8655].

[Pat96] J. Patarin. Hidden Fields Equations (HFE) and Isomorphisms of Polynomials (IP): Two New Families of Asymmetric Algorithms. In *Advances in cryptology – EUROCRYPT '96 (Saragossa, 1996)*, vol. 1070 of *Lecture Notes in Comput. Sci.*, pp. 33–48. Springer, Berlin, 1996.

[SPCK00] A. Shamir, J. Patarin, N. Courtois and A. Klimov. Efficient Algorithms for solving Overdefined Systems of Multivariate Polynomial Equations. In *Advances in cryptology – EUROCRYPT '00 (, 2000)*, vol. 1807 of *Lecture Notes in Computer Science*, pp. 392–407. Springer-Verlag, Heidelberg, 2000.

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Unité de recherche INRIA Futurs : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399